A STRUCTURED NONNEGATIVE MATRIX FACTORIZATION FOR SOURCE SEPARATION

Clément Laroche[⋆†] *Matthieu Kowalski*[†‡] *Hélène Papadopoulos*[†] *Gaël Richard*[⋆]

[⋆] Institut Mines-Telecom, Telecom ParisTech, CNRS-LTCI, Paris, France
[†] Univ Paris-Sud-CNRS-CentraleSupelec, L2S, Gif-sur-Yvette, France
[‡] Parietal project-team, INRIA, CEA-Saclay, France

ABSTRACT

In this paper, we propose a new unconstrained nonnegative matrix factorization method designed to utilize the multilayer structure of audio signals to improve the quality of the source separation. The tonal layer is sparse in frequency and temporally stable, while the transient layer is composed of short term broadband sounds. Our method has a part well suited for tonal extraction which decomposes the signals in sparse orthogonal components, while the transient part is represented by a regular nonnegative matrix factorization decomposition. Experiments on synthetic and real music data in a source separation context show that such decomposition is suitable for audio signal. Compared with three state-of-the-art harmonic/percussive decomposition algorithms, the proposed method shows competitive performances.

Index Terms— nonnegative matrix factorization, projective nonnegative matrix factorization, audio source separation, harmonic/percussive decomposition.

1. INTRODUCTION

Introduced by Lee & Seung [1], Non-Negative Matrix Factorization (NMF) has been widely used in a large variety of fields. In particular, this decomposition technique has been applied with great success in audio signal processing for automatic transcription [2, 3] and audio source separation [4, 5].

The goal of NMF is to approximate a data matrix $V \in \mathbb{R}_+^{n\times m}$ as:

$$V \approx \tilde{V} = WH \tag{1}$$

with $W \in \mathbb{R}_+^{n\times k}$ and $H \in \mathbb{R}_+^{k\times m}$ and where k is the rank of factorization, typically chosen such that $k(n+m) << nm$. As the data matrix V is usually redundant, the product WH can be thought as a compressed form of V, where W is a *dictionary* or a set of *patterns* and where H contains the *expansion coefficients*.

However, in practice, it is not guaranteed that the obtained decomposition has a valuable semantic interpretation. To alleviate this problem, it is necessary to exploit some prior information or to impose some constraints on the decomposition. For instance, information from the scores or from midi signals is used in a so-called supervised NMF in [2]. This method improves the automatic transcription accuracy, but requires well organized prior information. Another strategy is to rely on specific constraints deduced from the characteristics of the processed signals. For example, it is shown in [6] that enforcing temporal smoothness improves the physical meaning of the decomposition. Similarly in [7], Canadas & al. used four constraints in order to achieve a specific harmonic/percussive decomposition. In this case, the four hyper-parameters linked to the constraints need to be optimized and the best setting is often signal dependent.

Concurrently, other methods aim at underlining some mathematical properties of the decomposition, for example the orthogonality between the nonnegative basis functions (or patterns). Projective NMF (PNMF) and orthogonal NMF (ONMF) rely on this property. PNMF was used for image processing [8] and for feature extraction and clustering [9]. PNMF revealed interesting properties in practice: a higher efficiency for clustering than NMF [8] and the generation of a much sparser decomposition than NMF [9]. These intrinsic properties are particularly interesting for audio source separation as shown in [7].

The main advantage of these approaches compared to constrained NMF is that sparsity or orthogonality is obtained as intrinsic properties so they avoid a tedious and often unsatisfactory hyperparameter tuning stage. However, these approaches do not have a sufficient flexibility to properly represent the complexity of an audio scene composed of multiple and concurrent harmonic and percussive sources.

In this paper, we propose a new decomposition technique suitable for audio source separation that takes advantage of the sparse decomposition of PNMF but that allows a better representation of complex audio signals. More precisely, the initial nearly-orthogonal decomposition obtained by PNMF is extended by a non-orthogonal component that reveals to be particularly relevant to represent percussive or transient signals. The merit of this new method termed *Structured Projected Nonnegative Matrix Factorization (SPNMF)* is experimentally demonstrated on synthetic signals and on a specific

H. Papadopoulos is supported by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme
This work was supported by a grant from DIGITEO

task of percussive/harmonic components separation of real audio signals. It is important here to underline that our approach relies on an optimization problem without the need of any hyperparameter.

The paper is organized as follows. In Section 2, the Projective and Orthogonal NMF are described and compared on a theoretical basis. The SPNMF is then introduced in Section 3. We detail our experimental protocol and the results obtained on synthetic and real audio signals in Section 4. Finally, some conclusions are drawn in Section 5.

2. PROJECTIVE AND ORTHOGONAL NMF

2.1. Overview

The aim of PNMF is to find a non negative projection matrix $P \in \mathbb{R}_+^{n\times n}$ such that $V \approx \tilde{V} = PV$. In [10] Yuan & al. proposed to seek P as an approximative projection matrix under the form $P = WW^T$ with $W \in \mathbb{R}_+^{n\times k}$ with $k \leqslant n$. The PNMF problem reads :

$$\min_{W\geqslant 0} ||V - WW^TV||^2 \tag{2}$$

where $\|.\|^2$ is the squared Frobenius norm (Euclidean distance).

The ONMF [8] consists in solving the following problem:

$$\min_{W\geqslant 0, H\geqslant 0} ||V - WH||^2 \quad \text{s.t.} \quad W^TW = I_k \tag{3}$$

In this method, orthogonality between nonnegative basis functions is enforced during the optimization process. In practice, it seems that PNMF and ONMF lead to similar decompositions, as the W matrix estimated by PNMF is almost orthogonal (i.e., $\|W^TW - I_k\|^2$ is small). The links between PNMF, ONMF and the regular NMF are discussed in the next section.

2.2. On the equivalence between NMF, PNMF and ONMF

Using a squared Euclidean distance between the data matrix V and its approximation WH, the NMF problems reads:

$$\min_{W,H\geq 0} \|V - WH\|^2 ,$$

where PNMF (resp. ONMF) adds the constraint $H = W^TV$ (resp. $W^TW = I$). Let us assume that V admits an NMF decomposition without any errors, i.e., one can find W and H of rank k such that $V = WH$. Then, one can easily prove that necessarily $H = (W^TW)^{-1}W^TV$. Now, as demonstrated in [11], an invertible matrix is nonnegative if and only if it is a monomial matrix, that is, up to a scaling and permutation matrix, we necessarily have $W^TW = I$. This result allows one to state the following theorem:

Theorem 1 *Let $V \in \mathbb{R}_+^{n\times m}$. Let $W_{nmf}, W_{onmf}, W_{pnmf} \in \mathbb{R}_+^{n\times k}$ be solutions of the NMF, ONMF and PNMF respectively, with similar notations for the $H \in \mathbb{R}_+^{k\times m}$ matrix. Suppose that $k \leq \min(m,n)$ and rank(W) =rank$(H) = k$, and that $V = W_{nmf}H_{nmf}$. Moreover, suppose that $V \in \text{span}(W_{nmf})$. Then, up to a scaling and permutation matrix, we have:*

$$W_{nmf} = W_{onmf} = W_{pnmf}$$

and

$$H_{nmf} = H_{onmf} = H_{pnmf} = W_{nmf}^TV.$$

Or, equivalently, there exists a scaling and a permutation matrix such that:

$$W_{nmf}^TW_{nmf} = I \quad \textit{and} \quad H_{nmf} = W_{nmf}^TV.$$

In practice, the assumption $V = WH$ does not hold as soon as $k < \min(n,m)$, hence the motivation to introduce PNMF and ONMF. However, it is interesting to stress that the orthogonality of W is a requirement to obtain a true projector for H. Moreover, in practice, W_{pnmf} is "almost" orthogonal, leading to an "almost projection" for H_{pnmf} as motivated by the authors in [10]. This remark has encouraged us to build upon PNMF instead of ONMF in the next section.

3. STRUCTURED PROJECTIVE NMF (SPNMF)

3.1. Principle

As stated in [7], harmonic instruments have sparse basis functions whereas percussive instruments have much flatter spectra. As the columns of W are orthogonal, when two sources overlap in the Time-Frequency (TF) domain only one basis function will represent the mixture which is not adequate for efficient separation. To overcome this problem, we propose to add a standard NMF decomposition term to the PNMF. With a similar technique as in [4], we increase the rank of the PNMF. Let $k = k' + e$ with e being the number of additional components. We can expect that most of the harmonic components will be represented by the orthogonal part while the percussive ones will be in the regular NMF components. Let V be the magnitude spectrogram of the input data. The model is then given by

$$V \approx \tilde{V} = W_1H_1 + W_2H_2, \tag{4}$$

where W_1H_1 is the almost orthogonal part with rank k' and W_2H_2 are e regular NMF components. Following the same idea as in section 2.2, we obtain :

$$H_1 = W_1^T(V - W_2H_2), \quad \text{iff} \quad W_1^TW_1 = I\,. \tag{5}$$

We then propose the *structured projected NMF* cost function:

$$\min_{W_1,W_2,H_2\geq 0} ||V - W_1W_1^T(V - W_2H_2) - W_2H_2||^2. \tag{6}$$

As in [4], e is kept smaller than k'. The goal here is to focus most of the energy in the orthogonal part to benefit from the sparse decomposition property of PNMF.

3.2. Multiplicative update rules

Similarly to the regular PNMF, multiplicative update rules can be obtained from the cost function (6). The optimization of W_1 is straightforward, using similar technique as in [10, 12] which consists in splitting the gradient $\nabla F(W_1)$ in its positive $[\nabla F(W_1)]^+$ and negative parts $[\nabla F(W_1)]^-$. One obtains:

$$\begin{aligned}[\nabla F(W_1)]^+ = [4(VH_2^TW_2^T + W_2H_2V^T)\\ +2W_1W_1^T(VV^T + W_2H_2H_2^TW_2^T)\\ +2(VV^T + W_2H_2H_2^TW_2^T)W_1W_1^T]W_1 \end{aligned} \quad (7)$$

and:

$$\begin{aligned}[\nabla F(W_1)]^- = [4VV^T + 4W_2H_2H_2^TW_2^T\\ +2W_1W_1^T(VH_2^TW_2^T + W_2H_2V^T)\\ +2(VH_2^TW_2^T + W_2H_2V^T)W_1W_1^T]W_1. \end{aligned} \quad (8)$$

We are now able to write the multiplicative updates used for SPNMF as

$$W_1 \leftarrow W_1 \otimes \frac{[\nabla F(W_1)]^-}{[\nabla F(W_1)]^+}.$$

Similar expressions are obtained for W_2, H_2, but are omitted here for the sake of brevity.

4. EXPERIMENTAL VALIDATION

4.1. Experimental Protocol

We compare SPNMF with PNMF and regular NMF. The squared Euclidean distance with multiplicative update rules as stated in [1] is used for NMF. The three algorithms are initialized with the same random positive matrices $W_{ini} \in \mathbb{R}^{n\times k}$ and $H_{ini} \in \mathbb{R}_+^{k\times m}$. The rank of factorization k is the same for all methods. The k audio components are extracted by Wiener filtering and are then grouped into musical sources following the oracle method introduced by Virtanen in [6]. The Signal-to-Noise Ratio (SNR) is computed between the jth separated components $\tilde{x_j}$ and the mth original sources x_m as:

$$\text{SNR}(m,j) = \frac{\tilde{x_j}^2}{(\tilde{x_j} - x_m)^2}$$

The component j is assigned to the source m which leads to the highest SNR. The results are compared by means of the Signal-to-Distortion Ratio (SDR), the Signal-to-Interference Ratio (SIR) and the Signal-to-Artifact ratio (SAR) of each of the separated sources using the BSS Eval toolbox provided in [13].

4.2. Synthetic Tests

The test signal models a mix of harmonic and percussive instruments. The harmonic part is simulated by a sum of sine waves that overlap in time and in frequency. The first sum simulates a $C(2)$ with fundamental frequency $f_0 = 131$ Hz, the other one a $G(2)$ with $f_0 = 196$ Hz. To simulate the percussive part, every 1 s, we add 0.1 s of Gaussian white noise. The signal is 5 sec. long and the sampling rate is 4000 Hz. We compute the Short Time Fourier Transform (STFT) with a 512 sample-long (0.128 s) Hann analysis window and a 50% overlap. Here, $k' = 2$ and $e = 1$. The spectrogram of the signal is represented in Figure 1. As our input signal has three sources, we expect that one source will be represented by one component. Results are presented in Table 1 and on the whole, SPNMF outperforms NMF and PNMF on this synthetic signal. The NMF separates the three sources but, as shown in [6], the lack of temporal continuity degrades the extraction of the original signal from the mixture leading to poor SDR and SIR for the noise component. For the PNMF, because of the frequency overlap, the orthogonal components do not succeed to represent the noise correctly. The SIR is also low which indicates a poor separation of the sources. The SPNMF model supposes that the sources do not overlap in the TF domain. However, in this example, the SPNMF extracts the three components and performs better than the other methods.

		NMF	PNMF	SPNMF
SDR (dB)	C(2)	11.21	7.97	**13.44**
	G(2)	10.44	13.16	**16.79**
	Noise	2.37	0.11	**13.38**
SIR (dB)	C(2)	14.20	**17.67**	13.79
	G(2)	17.60	4.94	**18.53**
	Noise	7.82	0.45	**20.40**
SAR (dB)	C(2)	15.57	14.38	**24.90**
	G(2)	18.99	20.66	**21.64**
	Noise	3.69	10.42	**14.38**

Table 1: Source separation performance for the synthetic signal.

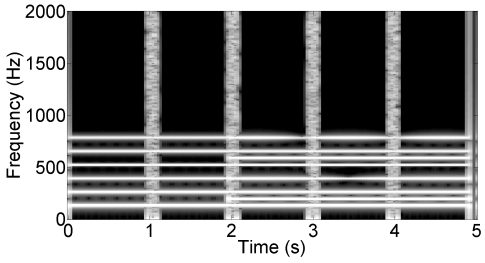

Fig. 1: Spectrogram of the synthetic test signals.

4.3. Tests on real audio signals

4.3.1. Database

The database is composed of monophonic real-world music excerpts. Each music signal contains percussive, harmonic instruments and vocals. The dataset is taken from SiSEC 2010 [14]. It consists of four recordings of duration ranging

from 14 to 24 s. To expand the tests, we add five extra audio files from the Medley-dB dataset [15]. The goal is to perform an harmonic/percussive decomposition as in [7] thus the vocal part is omitted. All the signals are sampled at $44.1kHz$. We compute the STFT with a 1024 sample-long Hann frame and a 50% overlap. Two tests are run on these data. The first test aims at comparing SPNMF, PNMF and regular NMF on the whole database. The second test aims at comparing SPNMF with three state-of-the-art methods on the SiSEC database.

4.3.2. *Evaluation of SPNMF, PNMF and NMF*

For the first test, the total number of components is set to $k = 16$ and the number of regular NMF components is set to $e = 5$. As in [4] the number of regular NMF components is set to a low value so that a maximum of the energy is in the orthogonal part. The results are displayed in figure 2. Each box-plot is made up of a central line indicating the median of the data, upper and lower box edges indicating the 1^{st} and 3^{rd} quartiles, and whiskers indicating the minimum and maximum values.

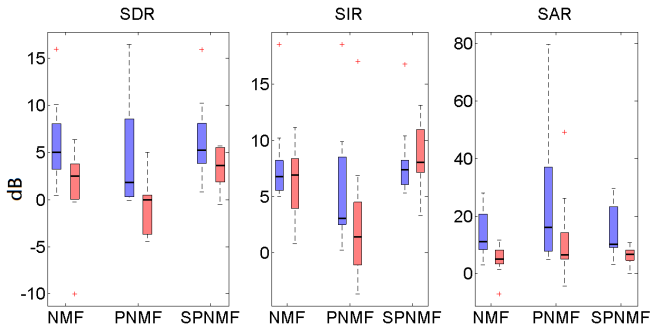

Fig. 2: SDR, SIR and SAR of harmonic (left bar)/percussive (right bar) estimated sources from the 9 test signals by NMF, PNMF and SPNMF.

Overall, the SPNMF outperforms NMF in terms of SDR and SIR. The SAR is similar for the two methods. These results on real audio signals confirm that SPNMF is more suitable than regular NMF to manage low interference source separation with TF overlap without having to optimize hyper-parameters.

The PNMF is outclassed by the other methods. This model is too restrictive for audio source separation because all the components tend to represent the harmonic instruments only. As a very small amount of energy is present in the percussive signal, it achieves very high scores for SAR and SIR.

4.3.3. *State-of-the-art benchmark*

For the second test, SPNMF is compared against three state of the art percussive/harmonic methods: HPSS [16], MFS [17], and constrained NMF [7]. These three algorithms are unsupervised.

Table 2 shows that our method outperforms the state of the art method on the first two songs, where the harmonic instruments have soft transients. The third song is a rap song with strong percussive sounds and a soft harmonic part. The SPNMF fails to separate the sources as they are blended in what the algorithm automatically attributes to the percussive part. Concurrently, the harmonic part has very low energy and does not obtain satisfying SDR and SIR scores. The TF overlap between the harmonic instruments can probably explain this drop in performance. On the last song the presence of prominent transients produced by the harmonic instruments degrade the performance. On this example the guitar and the snare drum are represented with the same component. The proposed method is outperformed on this song.

On average, SPNMF reaches a better percussive separation than MFS and HPSS, but obtains lower results for the harmonic separation. The initial sources together with the estimated sources of the experiments can be found on our web page[1].

4.3.4. *Discussion*

The results of section 4.2 show that having only an orthogonal basis (PNMF case) is too restrictive to reach a good quality of source separation. In the case of SPNMF, the tonal layer is extracted in the orthogonal portion and the noise is extracted in the unconstrained part, as expected from the model. The results of section 4.3 show that on real-world audio signals, SPNMF reaches state of the art results. The method is particularly efficient when pitched instruments have soft transients and when percussive sounds are not too prominent. In this case, the signal corresponds well to the model stated in (4) because the harmonic and the percussive parts have well defined different structures. However, if harmonic instruments have strong attacks, our SPNMF is outperformed by other state of the art methods because harmonic instruments are not well represented by the orthogonal basis functions.

5. CONCLUSION

In this paper, we demonstrated that SPNMF is a very promising model as an unconstrained decomposition method. Indeed, for synthetic and real audio data, the SPNMF is able to extract sources with less interference and better quality. On a harmonic/percussive separation task, it obtains similar results as purposely built methods. It can also retrieve a well structured decomposition with the tonal layer mostly extracted by the orthogonal part, while the transients are represented by the regular NMF components.

Future work on SPNMF will be dedicated to design of efficient initialization strategies. For instance, W_2 and H_2 can

[1] http://perso.telecom-paristech.fr/laroche/Article/EUSIPCO2015/

	HPSS [16]			MFS [17]			Constrained NMF [7]			SPNMF		
Percussive Separation	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
T2_01	2.6	**13.2**	1.1	-0.2	-1.5	**8.4**	4.0	6.5	5.7	**4.3**	11.0	5.6
T2_02	2.4	**10.2**	3.4	3.1	8.0	4.9	**5.2**	8.3	**7.5**	5.0	9.7	7.2
T2_03	2.6	**6.9**	4.0	2.5	2.1	**12.3**	**2.8**	2.6	11.1	1.5	6.6	4.0
T2_04	5.5	**11.5**	6.5	6.2	9.6	8.0	**7.5**	10.3	**10.3**	3.6	6.5	7.7
Mean	3.2	**10.5**	3.8	2.9	4.6	8.4	**4.9**	7.0	**8.7**	3.6	8.4	6.1
Harmonic Separation	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
T2_01	9.8	13.8	11.9	7.1	13.8	11.5	**11.0**	**14.8**	13.9	7.3	7.4	**28.3**
T2_02	4.8	6.3	9.8	5.5	**16.2**	11.6	**7.5**	9.3	12.1	6.5	7.8	**12.8**
T2_03	4.8	8.7	6.3	4.6	**11.0**	8.0	**5.0**	9.1	**8.6**	3.8	6.2	8.4
T2_04	5.6	**11.5**	6.7	6.2	9.3	8.7	**7.5**	10.6	**10.5**	4.7	6.6	10.1
Mean	6.3	10.1	8.7	5.9	**12.6**	10.0	**7.8**	11.0	11.3	5.6	7.0	**14.9**

Table 2: Harmonic/percussive source separation performance (in dB)

be forced to represent most of the transient part in order to obtain an fully unsupervised harmonic/percussive decomposition.

REFERENCES

[1] D. Lee and S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] S. Ewert and M. Müller, "Score-informed source separation for music signals," *Multimodal music processing*, vol. 3, pp. 73–94, 2012.

[3] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proc. of IEEE ICASSP*, 2007, vol. 1, pp. 65–68.

[4] R. Hennequin, R. Badeau, and B. David, "Time-dependent parametric and harmonic templates in non-negative matrix factorization," in *Proc. of DAFX*, 2010, pp. 246–253.

[5] J. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 564–575, 2010.

[6] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.

[7] F. Canadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, J. Carabias-Orti, and P. Cabanas-Molero, "Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–17, 2014.

[8] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *Proc. of IEEE IJCNN*, 2008, pp. 1828–1832.

[9] Z. Yang and E. Oja, "Linear and nonlinear projective nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 21, no. 5, pp. 734–749, 2010.

[10] Z. Yuan and E. Oja, "Projective nonnegative matrix factorization for image compression and feature extraction," *Image Analysis*, pp. 333–342, 2005.

[11] D. Richman and H. Schneider, "Primes in the semigroup of non-negative matrices," *Linear and multilinear algebra*, vol. 2, pp. 135–140, 1974.

[12] D. Lee and S. Seung, "Algorithms for non-negative matrix factorization," *Proc. of NIPS*, pp. 556–562, 2001.

[13] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, pp. 1462–1469, 2006.

[14] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Duong, "The 2010 signal separation evaluation campaign : audio source separation," in *Proc. of LVA/ICA*, 2010, pp. 114–122.

[15] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research," in *proc. of ISMIR*, 2014.

[16] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. of EUSIPCO*, 2008.

[17] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. of DAFX*, 2010.